4 Prisoner's Dilemma doesn't explain much

Robert Northcott and Anna Alexandrova

1. Introduction

The influence of the Prisoner's Dilemma on economics, law, political science, sociology, and even anthropology and biology is hard to overstate. According to JSTOR, almost 16,000 articles about it have appeared since 1960, with no sign of slowing down: 4,400 were just in the last 10 years. It has a high profile in non-academic media too. It appears as an explanation of phenomena as disparate as business strategy, political bargaining, gender relations and animal behavior. Historians of social science have referred to the Prisoner's Dilemma as a "mainstay" (Morgan 2012: 348) and an essential "set piece" (Rodgers 2011: 64). And according to Robert Axelrod, "the two-person iterated Prisoner's Dilemma is the *E. coli* of the social sciences" (quoted in McAdams 2009: 214)

As philosophers, our aim is to assess whether this development has been worthwhile and furthered the goals of social science. We ask this question even knowing that it cannot be answered fully in a single article. The research programs that the Prisoner's Dilemma has inspired are many and diverse, and the Prisoner's Dilemma is only one of many models that have been used in them. In addition, social science, like science in general, has many different goals and a judgment of worthwhileness requires a devilishly complex integration of conflicting considerations and values. Finally, sixty years may or may not be a sufficient span to judge. Nevertheless, we will brave giving a prima facie case that on at least one central criterion, namely providing causal explanations of field phenomena involving human co-operation, the Prisoner's Dilemma has failed to live up to its promise.

Before we start, two clarifications are in order. First, we do not wish to criticize the use of the Prisoner's Dilemma on moral or political grounds. It might be that teaching and using it makes people behave more selfishly and normalizes a narrow conception of rationality (Dupré 2001, Marwell and Ames 1981). But our concern is purely methodological: has the Prisoner's Dilemma delivered empirical success?

Second, we focus on the Prisoner's Dilemma because it is the subject of this volume, not because it is unique in the way it has been misused. Much of what we say applies to other analyses of collective action problems, and much of economic theory more generally. But here our focus will be on the Prisoner's Dilemma only. It is quite plausible that often the Prisoner's Dilemma gets misused just because it is uniquely famous, so scholars invoke it when instead they should be invoking a different game, say the Stag Hunt, or another co-ordination game (McAdams 2009). That is a mistake, but not the one we care to correct here, if only because correcting it would call for greater use of the very models that we argue do not provide a good return on investment anyway.

In Section 4.2 we present an account of how the Prisoner's Dilemma could provide causal explanations. The heart of the paper is in Section 4.3, where we make the case that in fact it has failed in this task. To this end, we examine in detail a famous purported example of Prisoner's Dilemma empirical success, namely Axelrod's analysis of WWI trench warfare, and argue that this success is greatly overstated. Further, we explain why that negative verdict is likely true generally, and not just in our case study. In Section 4.4, finally, we address some possible defenses of the Prisoner's Dilemma.

4.2 The possibility of explanation

4.2.1 What sort of explanation?

Is the Prisoner's Dilemma explanatory? There exists a canonical account of explanation known as *situational analysis* (Koertge 1975), which was originally articulated for social science by Popper, Dray and Hempel. As Mary Morgan, among others, has pointed out, the Prisoner's Dilemma is particularly well suited to it. According to situational analysis, social scientists do not seek laws as such but rather work to define "a kind or type of event" (Popper, quoted in Morgan 2012: 358). Such a type consists in certain features of a situation, which include the circumstances (institutional, historical, environmental) of agents, plus their beliefs and desires. As a second step, one adds an analysis of what it is rational to do in these particular circumstances. The third step is the assumption that the agents are indeed rational, and then the explanation follows: a given phenomenon arises because rational agents behave thus and thus in such and such situations. Since model building in game theory follows something like this logic, the claim is that situational analysis is how these models provide explanations. Theory building on this

view amounts to generating a portfolio of models which represent typical situations that arise in different domains of the social world. The Prisoner's Dilemma is one such model.

This leaves hanging an obvious question: exactly what sort of explanation does situational analysis provide? Accounts of scientific explanation abound. We will review here only the candidates most likely to apply to the Prisoner's Dilemma, without claiming any general superiority for one model of explanation over another.

If any theory of explanation can claim to be dominant in social science it is *causal* explanation. One well-known reason is its intimate connection to interventions, because interventions in turn are the lifeblood of policymaking. One prominent theory states that to give a causal explanation is to make a counterfactual claim that if a hypothetical intervention changed the cause then the effect would also be changed (Woodward 2003). We believe that something like this is the best hope for defenders of the explanatory potential of the Prisoner's Dilemma. But before returning to it, we will briefly mention two other leading possibilities.

The Prisoner's Dilemma in particular and game theory more generally is often thought to *unify* social phenomena: not just many different economic phenomena can be modeled but also political, legal, social, and personal ones too.[1] It is this unifying ambition that has earned economics more generally the accusation of imperialism. If the Prisoner's Dilemma really did unify phenomena in an explanatory way, we would welcome that and count it to the Prisoner's Dilemma's credit. But it does not. A closer look at unificationist theories of explanation, such as Kitcher's (1981), shows why. According to Kitcher, in order to explain, a theory must satisfy *two* unifying criteria: the first, roughly speaking, is range, i.e. the theory must indeed address explananda from many domains. But there is also a second criterion, which Kitcher calls *stringency*. Roughly speaking, this demands that such a unification not be vacuous – a theory must rule some things out, otherwise its compatibility with many domains is won too cheaply. Yet utility maximization, for instance, is underconstrained: utility is defined so thinly that almost anything could be an example of its maximization. This and other similar points tell against the claim that the Prisoner's Dilemma explains by unification.[2] Most likely, the needed

[1] And even sub-personal ones, as in Don Ross's game-theoretical approaches to the brain (Ross 2009).

[2] See Reiss (2012: 56–59) for more detail on why economic models do not satisfy unification theories of explanation.

constraints would have to come from causal knowledge about the contextual variation of judgment and choice, so causal explanation will return to the scene.

We believe that there is similarly no refuge in the notion of *mathematical* explanation. Perhaps, for instance, it might be thought that the Prisoner's Dilemma demonstrates the mathematical reason why two agents, given certain preferences and information and a certain environmental situation, will act in a particular way – much as statistical mechanics demonstrates the mathematical reason why with overwhelming probability heat will flow from hot air to cold. But, first, the notion of mathematical explanation of physical facts is contentious and the subject of much current debate.[3] And, second, in any case it is agreed by all that to be considered seriously mathematical explanations require empirical confirmation of precisely the kind that, we will argue, is typically absent in Prisoner's Dilemma cases.

Return now to situational analysis. This, we submit, can be thought of as an instance of causal explanation. When a social phenomenon is explained by the fact that it is an instance of a Prisoner's Dilemma, there is a claim to the effect that the structure of the situation in conjunction with the actor's rationality caused the outcome. This structure, the agents' beliefs and desires, and their rationality, are individually necessary and jointly sufficient causes of the outcome. Manipulating one of these conditions, say by changing the incentives or the information, would turn the situation into something other than a Prisoner's Dilemma and a different effect would obtain. To say that a situation is a Prisoner's Dilemma is thus just to specify a particular causal set-up.[4]

What causal structure does a Prisoner's Dilemma specify? According to Mary Morgan, there are three identity conditions for a Prisoner's Dilemma: (1) the 2-by-2 matrix which gives each player two options, (2) the inequalities that define the payoff structure, and (3) the narrative. The first two are well-known and uncontroversial, but the third ingredient is worth pausing on – what is a narrative and why would one think it essential to explanation?

As a story with a beginning, middle, and end, a narrative is the standard way of presenting the Prisoner's Dilemma. Originally at RAND, the story was

[3] See, for instance, recent work by Alan Baker, Bob Batterman, Chris Pincock, Mark Colyvan and Otavio Bueno. For an overview see Mancuso (2011).

[4] Admittedly, this claim runs afoul of the longstanding alternative according to which reason-based explanations cannot be causal because reasons have a normative connection to actions (e.g. Risjord 2005). If they cannot, then situational analysis is not a species of causal explanation after all. We do not wish to wade into this debate here, beyond saying that reason-based explanations routinely get re-cast as causal explanations by scientists and philosophers alike, and arguably for good reason.

Tosca's and Scarpia's attempt and failure to double-cross each other at the end of the opera *Tosca*. Later on, two prisoners' failure to cooperate against a prosecutor became the dominant story instead. Morgan insists that this storytelling aspect of situational analysis is essential but one that tends to get sidelined.[5] Yet in her view it makes the Prisoner's Dilemma what it is. First, a narrative matches the model and an actual situation – an explanandum – by providing a description of the model situation that the actual situation is supposed to match. It is thus a condition of model application. Second, a narrative provides a general category that allows for the classification of a situation as being of a particular *type*. Since situational analysis explains precisely by specifying a situation's type, the narrative is thus also essential to explanation (Morgan 2012: 362–363).

We think Morgan is right that the narrative does the explaining that the matrix and inequalities alone cannot. If she isn't right, then the whole story about situational analysis has to be abandoned too. A narrative tells us how informational and institutional constraints made agents behave as they did by providing reasons for them to behave as they did. If these constraints had been different, the agents would have behaved differently. So the narrative is essential to the explaining.[6,7]

An independent motivation for thinking that narratives are necessary for model-based explanation is provided by any view of economic models that does not take them to establish causal mechanisms. For example, on the open formula view of models, models by themselves do not amount to a causal claim but only to a template for such a claim that needs to be filled in using knowledge from outside of the model (Alexandrova 2008). This view is motivated by the numerous idealizations in economic models that cannot be relaxed and cannot be part of a causal mechanism that explains the target phenomenon. Accordingly, a model must instead be treated as merely a template or open formula. It is only the causal claim developed on the basis of the open formula that does the explaining – and is what Morgan calls the narrative.

[5] Here is Ken Binmore doing such sidelining: "Such stories are not to be taken too seriously. Their chief purpose is to serve as a reminder about who gets what payoff." (Binmore 1994: 102)

[6] The revealed preference approach would reject this account of how Prisoner's Dilemma causally explains, denying that we need or should appeal to reasons. (Binmore himself holds this view, which might explain his dismissal of the role of narratives.) We discuss this in Section 4.4.2 below.

[7] This leaves open how situational analysis could be extended to cases where the actors in a Prisoner's Dilemma are not individual humans. We do not discuss that issue here, except to note we don't think there is any a priori reason why it couldn't be.

Accepting for now that this is how the Prisoner's Dilemma could explain, we move on to a potential obstacle. How could the Prisoner's Dilemma explain given that it is so *idealized*?

4.2.2 Prisoner's Dilemma and idealization

By the standards of other models in microeconomics, the Prisoner's Dilemma is remarkably undemanding. The simplest version can be formulated with only an ordinal utility function, not a cardinal one. As a result it needs only the minimal consistency axioms on preferences (completeness, transitivity and asymmetry) and not the more controversial rankings of lotteries that the von Neumann-Morgenstern expected utility maximization framework requires. In addition to this the single-shot equilibrium, i.e. defection by both players, can be justified by dominance analysis only. It is thus not necessary to assume that players follow Nash equilibrium and hence co-ordinate their beliefs about each other. In this sense the Prisoner's Dilemma relies on far fewer controversial assumptions than do other models in game theory.

But it is still idealized nevertheless. It postulates an invariable choice of a dominant strategy by perfectly consistent agents. Actual people are not like this, as many experiments show, and that is already enough to query how such a model (or model plus narrative) could be explanatory. Can idealization ever be reconciled with explanation? Most certainly it can. Philosophers of science have come up with various accounts to make sense of the widespread explanatory use of seemingly false models.[8] We do not need to go into the details here. Roughly, they all come down to the same verdict: idealized models can be explanatory in the causal sense when their falsity does not matter, i.e. when the idealizations are true enough for the purposes at hand.

But for the Prisoner's Dilemma this defense will generally not work. Evidence from behavioral economics about how deeply context affects judgment and choice is robust. And social situations that approximate the single-shot or iterated Prisoner's Dilemma either in the field or in the laboratory exhibit a great deal of variability in levels of co-operation, enough to raise questions about the Prisoner's Dilemma's predictive value. Nevertheless, this still leaves open the possibility that the Prisoner's Dilemma does work in a few important cases. We turn to that issue now.

[8] See, for instance, recent work by Nancy Cartwright, Daniel Hausman, Uskali Mäki, Michael Strevens, and Michael Weisberg. For an overview see Weisberg (2012).

4.3 The reality of non-explanation

4.3.1 Casual empiricism

Various encyclopedia entries and overview articles across economics and philosophy discuss some of the Prisoner's Dilemma literature's main developments: asymmetric versions, versions with multiple moves or players, single-person interpretations, versions with asynchronous moves, finitely and infinitely and indefinitely iterated versions, iterated versions with error, evolutionary versions, versions interpreted spatially, and many other tweaks besides (Govindan and Wilson 2008, Michihiro 2008, Kuhn 2009). Many of these are apparently motivated by a loose kind of responsiveness to real-world problems and complications. After all, putative actual players of the Prisoner's Dilemmas will often act asynchronously or more than once, or make errors, and so on. Certainly, the subtlety and sophistication of this work is often impressive. Nevertheless, a striking fact about it is its overwhelmingly *theoretical* focus. The underlying motivation by real-world concerns is somewhat casual. Deeper empirical applications of the Prisoner's Dilemma, featuring detailed examination of the evidence of particular real-world cases, are remarkably thin on the ground.

The overall picture is that research muscle has been bet on theoretical development rather than empirical applications.[9] It is in fact hard to find serious attempts at applying the Prisoner's Dilemma to explain actual historical or contemporary phenomena. We have found that the instances in which the Prisoner's Dilemma is mentioned in empirical contexts tend to come in two kinds. The first kind are the purely informal mentions in textbooks, blog posts, teaching tools or offhand remarks in the media of the sort: "Well, that's obviously a Prisoner's Dilemma!"[10] Clearly, merely identifying a casual similarity between the Prisoner's Dilemma and an actual situation does not count as explanatory success. Sure, the price war between two gas stations may look like a Prisoner's Dilemma in some respects, but in other respects it doesn't. It would need to be explained why the dissimilarities do not matter.

The second kind of empirical use is far from casual. Ever since the discovery and proliferation of game theory in Cold War US academia, a great many fields in social science have adopted the language of the Prisoner's Dilemma (among other models) to reconceive old explananda, be they in industrial

[9] A lot of the Prisoner's Dilemma literature is "empirical" in the sense that it reports on psychological experiments. We discuss these in Section 4.4.1 below.

[10] http://cheaptalk.org/2013/11/13/prisoners-dilemma-everywhere-amazon-source/

organization or international bargaining (Erickson et al 2013, Jervis 1978). But again, only rarely are game theory models applied carefully to specific field phenomena, and when they are it is not the Prisoner's Dilemma that is used. For the most part, the game theory models instead play a research-structuring rather than explanatory role, defining an agenda for the disciplines in question (see also Section 4.4.3).

4.3.2 A case study: Prisoner's Dilemma and World War I truces

Surveying the social sciences one finds a great many instances where the Prisoner's Dilemma is mentioned as explaining a field phenomenon. But the closer one looks, the more elusive explanatory success becomes. In the limited space here, we will support this claim via an extended analysis of one example. Of course, a single case does not prove much by itself. But if the Prisoner's Dilemma's explanatory shortcomings only become apparent when one looks at the fine details, then it is much more instructive to look at one case in depth than at many cases superficially.

The particular case we will examine is the "live-and-let-live" system that arose in World War I (WWI) trenches, which Robert Axelrod analyzed in terms of the Prisoner's Dilemma in chapter 4 of his book (1984). It is the most famous example of a detailed application of the Prisoner's Dilemma to a particular real-world target. It is also arguably the best one too, even though the details of Axelrod's analysis have subsequently been challenged (see Section 4.3.3 below).

Axelrod draws on the fascinating and detailed account of WWI trench warfare by the historian John Ashworth (1980), itself based on extensive letters, archives, and interviews with veterans. The "live-and-let-live" system refers to the many informal truces that arose on the Western front. "Truces" here covers complete non-aggression, temporary periods of non-aggression (e.g. at mealtimes), certain areas of non-aggression (e.g. mutually recognized "safe areas"), or many other mutual limitations on aggression (e.g. intricate local norms covering what actions and responses were or were not "acceptable"). The striking fact is that such truces between enemies arose spontaneously despite constant severe pressure against them from senior commanders. How could this have happened?

Axelrod's case is that, upon analysis, the implicit payoffs for each side on the front formed a Prisoner's Dilemma, and that this is an excellent example of how the Prisoner's Dilemma can illuminate a real-world phenomenon. In particular, he argues that the situation was an indefinitely iterated Prisoner's

Dilemma, and that co-operation – i.e. a truce – was therefore exactly the Prisoner's Dilemma's prediction.[11]

Axelrod is quite explicit that his goal is explanation, and of multiple explananda (1984: 71):

The main goal [of the WWI case study] is to use the theory to explain:

1) How could the live-and-let-live system have gotten started?
2) How was it sustained?
3) Why did it break down toward the end of the war?
4) Why was it characteristic of trench warfare in World War I, but of few other wars?

A second goal is to use the historical case to suggest how the original concepts and theory can be further elaborated.

Of course, he is well aware of the many real-life complications. But he defends the application of the Prisoner's Dilemma nevertheless (1984: 19): "The value of an analysis without [the real-life complications] is that it can help to clarify some of the subtle features. . . which might otherwise be lost in the maze of complexities of the highly particular circumstances in which choices must actually be made. It is the very complexity of reality which makes the analysis of an abstract interaction so helpful as an aid to understanding."

Axelrod's meaning is less clear here, but perhaps his aims can be interpreted as some combination of explanation, heuristic value, and understanding, and maybe also the unificatory virtue of generalizability across contexts. Certainly, these seem very reasonable goals. Indeed, if applying the Prisoner's Dilemma did not achieve any of these, what would be the gain from applying it at all? So let us examine how well Axelrod's study fares by these criteria.

Many historical details do seem to tell in its favor:

- Breaches of a truce were followed by retaliation – but only on a limited scale. This is consistent with Tit-for-Tat.

[11] In fact, of course, the indefinitely iterated Prisoner's Dilemma has many other Nash equilibria besides mutual cooperation. The analysis that Axelrod actually applies comes from his well-known Prisoner's Dilemma computer tournaments, the winner of which he concluded was the Tit-for-Tat strategy with initial cooperation (Section 4.3.3). If adopted by both players, this strategy predicts indefinite mutual co-operation. Throughout this section, we will use "Prisoner's Dilemma" as shorthand for this richer theoretical analysis of Axelrod's. (The main lesson, namely the difficulty of establishing the Prisoner's Dilemma's explanatory success, would apply still more strongly to the Prisoner's Dilemma alone, because then we would be faced with the additional problem of equilibrium selection too.)

- Both sides often demonstrated their force capability – but in harmless ways, such as by expertly shooting up a harmless barn. Axelrod argues that Tit-for-Tat predicts that a credible threat is important to making co-operation optimal, but that actually defecting is not optimal. Hence, ways of establishing credibility in a non-harmful manner are to be expected.
- The Prisoner's Dilemma predicts that iteration is crucial to maintaining a truce. Soldiers actively sought to ensure the required continuity on each side, even though individual units were often rotated. For instance, old hands typically instructed newcomers carefully as to the details of the local truce's norms, so that those norms often greatly outlasted the time any individual soldier spent on that front.

Perhaps Axelrod's most striking evidence is how the live-and-let-live system eventually broke down. The (unknowing) cause of this, he argues, was the beginning of a policy, dictated by senior command, of frequent *raids*. These were carefully prepared attacks on enemy trenches. If successful, prisoners would be taken; if not, casualties would be proof of the attempt. As Axelrod observes:

> There was no effective way to pretend that a raid had been undertaken when it had not. And there was no effective way to co-operate with the enemy in a raid because neither live soldiers nor dead bodies could be exchanged. The live-and-let-system could not cope with the disruption. . . since raids could be ordered and monitored from headquarters, the magnitude of the retaliatory raid could also be controlled, preventing a dampening of the process. The battalions were forced to mount real attacks on the enemy, the retaliation was undampened, and the process echoed out of control. (Axelrod 1984: 82)

The conditions that the Prisoner's Dilemma predicts as necessary for co-operation were unavoidably disrupted and, Axelrod argues, it is no coincidence that exactly then the truces disappeared.

We agree that many of the historical details are indeed, in Axelrod's phrase, "consistent with" the situation being an iterated Prisoner's Dilemma.[12] Nevertheless, upon closer inspection, we do not think the case yields any predictive

[12] As we will see, many other details were *not* so consistent. But even if they all had been, this criterion is far too weak for explanation. After all, presumably the WW1 details are all consistent with the law of gravity too, but that does not render gravity explanatory of them.

or explanatory vindication of the Prisoner's Dilemma, contrary both to Axelrod's account and to how that account has been widely reported.

Why this negative verdict? To begin, by Axelrod's own admission some elements of the story deviate from his Prisoner's Dilemma predictions. First, the norms of most truces were not Tit-for-Tat but more like Three-Tits-for-Tat. That is, retaliation for the breach of a truce was typically three times stronger than the original breach.[13] Second, in practice two vital elements to sustaining the truces were the development of what Axelrod terms ethics and rituals: local truce norms became ritualized, and their observance quickly acquired a moral tinge in the eyes of soldiers. Both of these developments made truces much more robust and are crucial to explaining those truces' persistence, as Axelrod concedes. Yet, as Axelrod also concedes, the Prisoner's Dilemma says nothing about either. Indeed, he comments (1984: 85) that this emergence of ethics would most easily be modeled game-theoretically as a change in the players' payoffs, i.e. potentially as a different game altogether.

Moreover, there are several other predictive shortfalls in addition to those remarked by Axelrod. First, Tit-for-Tat predicts that there should be no truce-breaches at all. Again, this prediction is incorrect: breaches were common. Second, as a result (and as Axelrod acknowledges), a series of dampening mechanisms therefore had to be developed in order to defuse post-breach cycles of retaliation. Again, the Tit-for-Tat analysis is silent about this vital element for sustaining the truces. Third, it is not just that truces had to be robust against continuous minor breaches; the bigger story is that often no truces arose at all. Indeed, Ashworth examined regimental and other archives in some detail to arrive at the estimate that, overall, truces existed about one-quarter of the time (1980: 171–175). That is, on average, three-quarters of the front was *not* in a condition of live-and-let-live. Again, the Prisoner's Dilemma is utterly silent as to why. Yet part of explaining why there were truces is surely also an account of the difference from those cases where there were *not* truces.[14]

Moreover again, the Prisoner's Dilemma does not fully address two other, related issues. The first is how truces originated as opposed to how they

[13] The Prisoner's Dilemma itself (as opposed to Tit-for-Tat) is silent about the expected level of retaliation, so should stand accused here merely of omission rather than error.

[14] Ashworth, by contrast, does develop a detailed explanation, largely in terms of the distinction between elite and non-elite units, and their evolving roles in the war. The escalation in the use of raids, so emphasized by Axelrod, is only one part of this wider story. Most areas of the front were not in a state of truce even before this escalation.

persisted, about which it is again completely silent.[15] The second is how truces ended. This the Prisoner's Dilemma does partly address, via Axelrod's discussion of raids. But many truces broke down for other reasons too. Ashworth devotes most of his chapter 7 to a discussion of the intra-army dynamics, especially between frontline and other troops, which were often the underlying cause of these breakdowns.

And moreover once more, Ashworth analyses several examples of strategic sophistication that were important to the maintenance of truces but that are not mentioned by Axelrod. One such example is the use by infantry of gunners. In particular, gunners were persuaded to shell opposing infantry in response to opponents' shelling, so that opposing infantry would then pressurize their own gunners to stop. This was a more effective tactic for reducing opponents' shelling than any direct attack on hard-to-reach opposing gunners (168). Another example: the details of how increased tunnelling beneath enemy trenches also disrupted truces, quite separately from increased raiding (199–202). Perhaps Axelrod's analysis could be extended to these other phenomena too; but in lieu of that, the Prisoner's Dilemma's explanatory reach here seems limited.

We have not yet even mentioned more traditional worries about rational choice explanations. An obvious one here is that the explanations are after-the-fact; there are no novel predictions. Thus it is difficult to rule out wishful after-the-fact rationalization, or that other game structures might fit the evidence just as well. A second worry is that Axelrod's crucial arguments that the payoff structure fits that of an iterated Prisoner's Dilemma are rather brief and informal (1984: 75). Do his estimations here really convince?[16] And are the other assumptions of the Prisoner's Dilemma, such as perfectly rational players and perfect information, satisfied sufficiently well?

In light of these multiple shortfalls, how can it be claimed that the Prisoner's Dilemma explains the WWI truces? It is not empirically adequate, and it misses crucial elements even in those areas where at face value it is empirically adequate. Moreover, it is silent on obvious related explananda, some of them cited as targets by Axelrod himself: not just why truces persisted but also why they occurred on some occasions but not on others, how they originated, and (to some degree) when and why they broke down.

[15] Again, Ashworth covers this in detail (as Axelrod does report).

[16] Gowa (1986) and Gelman (2008), for instance, argue that they do not. (Gowa also voices some of our concerns about the explanatory adequacy of Axelrod's analysis as compared to Ashworth's.)

But note that there is no mystery as to what the actual causal explanations of these various explananda are, for they are given clearly by Ashworth and indeed in many cases are explicit in the letters of the original soldiers. Thus, for instance, elite and non-elite units had different attitudes and incentives, for various well-understood reasons. These in turn led to truces occurring overwhelmingly only between non-elite units, again for well-understood reasons. The basic logic of reciprocity that the Prisoner's Dilemma focuses on, meanwhile, is ubiquitously taken by both Ashworth and the original soldiers to be so obvious as to be mentioned only briefly or else simply assumed. Next, why did breaches of truces occur frequently, even before raiding became widespread? Ashworth explains via detailed reference to different incentives for different units (artillery versus frontline infantry, for instance), and to the fallibility of the mechanisms in place for controlling individual hotheads (1980: 153–171). And so on. Removing our Prisoner's Dilemma lens, we see that we have perfectly adequate explanations already.

Overall, we therefore judge both that the Prisoner's Dilemma does not explain the WWI truces, and that we already have an alternative – namely, historical analysis – that does. So if not explanation, what else might the Prisoner's Dilemma offer? What fallback options are available? It seems to us there are two. The first is that, explanatory failure notwithstanding, the Prisoner's Dilemma nevertheless does provide a deeper "insight" or "understanding," at least into the specific issue of why the logic of reciprocity sustains truces. We address this response elsewhere (Northcott and Alexandrova 2013). In brief, we argue that such insight is of no independent value without explanation, except perhaps for heuristic purposes.

This leads to the second fallback position – that even if the Prisoner's Dilemma does not provide explanations here, still it is of heuristic value (see also Section 4.4.3 below). In particular, presumably, it is claimed to guide us to those strategic elements that do provide explanation. So does the Prisoner's Dilemma indeed add value in this way to our analysis of the WWI truces? Alas, the details suggest not, for two reasons.

First, the Prisoner's Dilemma did not lead to any causal explanations that we didn't have already. To see this, one must note a curious dialectical ju-jitsu here. Axelrod cites many examples of soldiers' words and actions that seem to illustrate them thinking and acting in Prisoner's Dilemma-like patterns. These are used to support the claim that the Prisoner's Dilemma is explanatory. (This is a common move in casual applications of the Prisoner's Dilemma more generally.) Yet now, having abandoned the explanatory claim and considering instead whether the Prisoner's Dilemma might be valuable

heuristically, these very same examples become evidence *against* its value rather than for it. This is because they now show that Prisoner's Dilemma-like thinking was present already. Ubiquitous quotations in Ashworth, many cited by Axelrod himself, show that soldiers were very well aware of the basic strategic logic of reciprocity. They were also well aware of the importance of a credible threat for deterring breaches (Ashworth 1980, 150). And well aware too of why frequent raiding rendered truces impossible to sustain, an outcome indeed that many ruefully anticipated even before the policy was implemented (Ashworth 1980: 191–198).[17]

The second reason why the Prisoner's Dilemma lacks heuristic value is that it actively diverts attention *away* from aspects that are important. We have in mind many of the crucial features already mentioned: how truces originated, the causes and management of the continuous small breaches of them, the importance of ethics and ritualization to their maintenance independent of strategic considerations, why truces occurred in some sections of the front but not in a majority of them, and so on.[18] Understanding exactly these features is crucial if our aim is to encourage co-operation in other contexts too – and this wider aim is the headline one of Axelrod's book[19], and implicitly surely a

[17] Ashworth reports (1980: 197): "One trench fighter wrote a short tale where special circumstances . . . [enabled the truce system to survive raids]. The story starts with British and Germans living in peace, when the British high command wants a corpse or prisoners for identification and orders a raid. The British soldiers are dismayed and one visits the Germans taking a pet German dog, which had strayed into British trenches. He attempts to persuade a German to volunteer as a prisoner, offering money and dog in exchange. The Germans naturally refuse; but they appreciate the common predicament, and propose that if the British call off the raid, they could have the newly dead body of a German soldier, providing he would be given a decent burial. The exchange was concluded; the raid officially occurred; high command got the body; and all parties were satisfied. All this is fiction, however. . ."

This soldier's fictional tale demonstrates vividly a very clear understanding of the Prisoner's Dilemma's strategic insights *avant la lettre*, indeed a rather more nuanced and detailed understanding than the Prisoner's Dilemma's own. No need for heuristic aid here.

[18] For example, a full understanding of why raiding disrupted truces goes beyond the simple Prisoner's Dilemma story. Ashworth summarises (1980: 198): "Raiding . . . replaced a background expectancy of trust with one of mistrust, making problematic the communication of peace motives; raids could not be ritualised; the nature of raids precluded any basis for exchange among adversaries; and raiding mobilised aggression otherwise controlled by informal cliques."

[19] Axelrod summarizes (1984: 21–22) the wider lessons of the WWI case for cooperation in this way: it can emerge spontaneously, even in the face of official disapproval; it can be tacit rather than explicit; it requires iterated interaction; and it does not require friendship between the two parties. But all these lessons are already contained in Ashworth's historical account – and, we argue, Ashworth establishes them rather better.

major motivation for the Prisoner's Dilemma literature as a whole. Yet here, to repeat, the Prisoner's Dilemma directs our attention away from them!

Overall, in the WWI case:

1) The Prisoner's Dilemma is not explanatory.
2) The Prisoner's Dilemma is not even valuable heuristically. Rather, detailed historical research offered much greater heuristic value, as well as much greater explanatory value.

Thus, Axelrod's own stated goals were not achieved. More generally, if this case is indicative then we should conclude that, at least if our currency is causal explanations and predictions of real-world phenomena, the huge intellectual investment in the Prisoner's Dilemma has not been justified.

4.3.3 It's not just Axelrod

Axelrod's work was innovative in that he arrived at his endorsement of Tit-for-Tat via a simulation rather than by calculation. For this reason, he has been credited with helping to kick-start the research program of evolutionary game theory. His engaging presentation also quickly won a popular following. Nevertheless, even theorists sympathetic to the potential of the Prisoner's Dilemma to explain cooperation have since then largely rejected the details of his analysis –not on the empirical grounds that we have emphasized, but rather on theoretical grounds. In particular, other simulations have not reproduced Tit-for-Tat's superiority; indeed, often "nasty" strategies are favored instead (e.g. Linster 1992). More generally, Axelrod's approach arguably suffers badly from a lack of connection to mainstream evolutionary game theory (Binmore 2001). The conclusion is that it is dubious that the WWI soldiers should be predicted to play Tit-for-Tat at all.

It does not follow, however, that Axelrod is therefore a misleadingly easy target – for two reasons. First, no better analysis of the WWI case has appeared. What strategy does best model soldiers' behavior in the trenches? This is neither known, nor has anyone bothered to find out. It is true that there are now much more sophisticated results from simulations of iterated Prisoner's Dilemma in different environments and, thus, better theoretical foundations. But there has been no attempt to use these improved foundations to model the WWI live-and-let-live system. Until a successor analysis has actually been applied to the WWI case, we have no reason to think it would explain the behavior in the trenches any better than did Axelrod's, let alone better than Ashworth does.

Second, it is not just that the WWI case in particular has been left ignored by the emphasis on theory. Rather, it is that the same is true of field cases generally. Detailed empirical engagement is very rare.[20] Of course, short of an exhaustive survey it is hard to prove a negative thesis such as this, but we do not think the thesis is implausible. One initial piece of evidence is that Axelrod's WWI study continues to be used in many textbooks as a prime example of the Prisoner's Dilemma's supposed explanatory relevance.[21] Perhaps these textbooks' selections are just ill judged, but the point is the perceived lack of alternative candidates.

Or consider the career of the Prisoner's Dilemma in biology – a discipline often cited by game theorists as fertile ground for applications. But the details turn out to be discouraging there too, and for a familiar reason, namely a focus on theoretical development rather than on field investigations:

> [T]he preoccupation with new and improved strategies has sometimes distracted from the main point: explaining animal cooperation . . . Understanding the ambiguities surrounding the Iterated Prisoner's Dilemma has stimulated 14 years of ingenious biological theorizing. Yet despite this display of theoretical competence, there is no empirical evidence of non-kin cooperation in a situation, natural or contrived, where the payoffs are known to conform to a Prisoner's Dilemma. (Clements and Stephens 1995)

And for a similarly negative verdict:

> [D]espite the voluminous literature, examples of Prisoner's Dilemma in nature are virtually non-existent. . . Certainly, with all the intense research and enthusiastic application of [the Prisoner's Dilemma] to real world situations, we may expect that we should have observed more convincing empirical support by now if it ever were to hold as a paradigm. . . (Johnson et al. 2002)

Payoff structures in field cases rarely seem to match those of the Prisoner's Dilemma, often because of the different values put on a given outcome by different players. Johnson et al. (2002) explain why several much-reported successes are in fact only dubiously cases of Prisoner's Dilemma at all, such as predator "inspection" in shoaling fish, animals cooperating to remove

[20] Sunstein (2007) comes close, but even here the phenomenon in question (the failure of the Kyoto Protocol) is explained in part by the fact that it does *not* have a Prisoner's Dilemma structure.

[21] E.g. Besanko and Braeutigam (2010: 587–588) – and there are many other examples.

parasites from each other, or lions cooperating to defend territory. The one exception they allow is Turner and Chao's (1999) study of an RNA virus. Even the game theorists Nowak and Sigmund (1999: 367), while lionizing the Turner and Chao case, concede that other claimed cases of the Prisoner's Dilemma occurring in nature are unproven. They also concede that, with reference to the literature in general, "it proved much easier to do [computer] simulations, and the empirical evidence lagged sadly behind."

Nor does there seem good reason to expect a dramatically different story in other disciplines. Gowa (1986), for instance, in a review of Axelrod's 1984 book, is generally sympathetic to the application of formal modeling. Nevertheless, she argues that the simple Prisoner's Dilemma template is unlikely to be a useful tool for studying the complex reality of international relations. And indeed since then bargaining models have become the norm in IR, because they can be purpose-built to model specific cases of negotiations in a way that the Prisoner's Dilemma can't be (e.g. Schultz 2001).

Overall, the Axelrod WWI case is therefore not a misleadingly soft target amid a sea of many tougher ones. On the contrary, it remains by far the most famous detailed application of the Prisoner's Dilemma to a field case for good reason – there aren't many others.

4.4 Defenses of Prisoner's Dilemma

4.4.1 Laboratory experiments

As we have noted, a large portion of the Prisoner's Dilemma literature concerns theoretical development, in which we include the running of the dynamics of idealized systems. Very little concerns close empirical analysis of field phenomena. But there is a third category that, although it is hard to quantify precisely, in terms of sheer number of papers might form the largest portion of all. This third category concerns psychology experiments, in particular simulation in the laboratory of the Prisoner's Dilemma or closely related strategic situations. Do the human subjects' actions in the laboratory accord with the predictions of theory? What factors are those actions sensitive to? Even a cursory sampling of the literature quickly reveals many candidates. For example, how much is cooperation in a laboratory setting made more likely if we use labeling cues (Zhong et al. 2007), if we vary payoffs asymmetrically (Ahn et al. 2007), if there is a prior friendship between players (Majolo et al. 2006), if players have an empathetic personality type (Sautter et al. 2007), or if players expect cooperation from opponents (Acevedo and Krueger

2005)? Literally thousands of articles are similar. Do they demonstrate, as it were, an empirical wing to the Prisoner's Dilemma literature after all? Unfortunately we think not, or at least not in the right way. Here are two reasons for this negative verdict.

First, the emphasis in most of this literature is on how a formal Prisoner's Dilemma analysis needs to be *supplemented*.[22] Typically, what makes cooperation more likely is investigated by manipulating things external to the Prisoner's Dilemma itself, such as the psychological and social factors mentioned above. That is, the focus of the literature is on how the Prisoner's Dilemma's predictions break down and on how instead a richer account, sensitive to otherwise unmodeled contextual factors, is necessary to improve predictive success. This is just the same lesson as from the WWI case – only now this lesson also holds good even in the highly controlled confines of the psychology laboratory.

Second, an entirely different worry is perhaps even more significant: whatever the Prisoner's Dilemma's success or otherwise in the laboratory, what ultimately matters most is its success with respect to *field* phenomena. Does it predict or explain the behavior of banks, firms, consumers, and soldiers outside the laboratory? Surely, that must be the main motivation for social scientists to use the Prisoner's Dilemma. Accordingly, the main value of the psychology findings, at least for non-psychologists, must be *instrumental* – are they useful guides to field situations? Presumably, they would indeed be if the psychological patterns revealed in experiments carried over reliably to field cases. Suffice to say here that such extrapolation is far from automatic, given the huge range of new contextual cues and inputs to be expected whenever moving from the laboratory to the field. The issue is the classic one of external validity, on which there is a large literature.[23] So far, the field evidence for the Prisoner's Dilemma is not encouraging.

[22] As Binmore and Shaked (2010) and others argue, other empirical work shows that, after a period of learning, the great majority of laboratory subjects do eventually defect in one-shot Prisoner's Dilemma games, just as theory predicts. Nevertheless it is uncontroversial that, initially at least, many or even most do not. It is this that has spawned the large literature investigating what contextual factors influence such instances of cooperation.

[23] Levitt and List (2007) discuss this from an economist's perspective with regard to cooperation specifically. Like everyone else, they conclude that external validity can rarely if ever be assumed. This is true even of field explananda that one might think especially close to laboratory conditions and thus especially promising candidates, such as highly rule-confined situations in TV game shows (see, e.g., van den Assem et al. 2012 about the Split or Steal show).

4.4.2 Revealed preferences to the rescue?

There is another way to defend the Prisoner's Dilemma's explanatory power. According to it, the Prisoner's Dilemma is not supposed to furnish explanations in which people co-operate because they *feel* it would be better for them and they can *trust* the other party to reciprocate; or fail to cooperate because they are *afraid* of being taken for a ride. Although these are the conventional articulations of what happens in a Prisoner's Dilemma, they are causal claims made using psychological categories such as feelings, judgments, and fears. They assume that behavior stems in part from these inner mental states and can be explained by them.

But a long tradition in economics maintains that this is exactly the wrong way to read rational choice models. Agents in these models do not make choices because they judge them to be rational; rather, the models are not psychological at all. To have a preference for one option over another just *is* to choose the one option when the other is available. This is the well-known revealed preference framework. It *defines* preferences as choices (or hypothetical choices), thus enforcing that economic models be interpreted purely as models that relate observable behavior to (some) observable characteristics of social situations.[24] On this view, agents cooperate in the Prisoner's Dilemma not because they feel they can trust each other, but rather because this is a game with an indefinite horizon in which the payoffs are such that rational agents cooperate. Although such an explanation sounds tautologous, it isn't. It comes with relevant counterfactual claims, such as that (given their history of choices) agents would not have cooperated if the game had been single-shot rather than iterated. This is a causal counterfactual and thus can be used for causal explanation. It only sounds tautologous because we are used to the natural and deeper psychological reading of the Prisoner's Dilemma in line with standard explanations of actions. But the revealed preference reading is perfectly conceivable too, and moreover the party line in economics is that it is in fact the correct one.

We will not discuss why the revealed preference view became popular within economics, nor evaluate whether it is viable in general.[25] Rather, our interest here is whether even according to it the Prisoner's Dilemma is a promising research program for explaining actual field cases. On this latter issue, we make two pessimistic points.

[24] Only "some" because, on the revealed preference view, data on what agents say, or on their physiological and neurological properties, are typically not deemed admissible even though they are perfectly observable.

[25] For up-to-date interpretations, criticisms, defenses, and references, see Hausman (2012).

First, a strict revealed preference theory of explanation seems needlessly philistine. To the extent that we have a good explanation for the live-and-let-live system in the WWI trenches it is in part a psychological explanation deeply steeped in categories such as fear, credibility, and trust. This is a general feature of social explanations – they are explanations that appeal to beliefs and desires (Elster 2007). For the revealed preference theorist, this is reason to dump them. But Ashworth's WWI explanations would be greatly impoverished if we did. In fact, not much of his rich and masterful analysis would remain at all.

Second, even if interpreted in revealed preference terms, the Prisoner's Dilemma would still state false counterfactual (or actual) claims. Many more factors affect behavior than just the ones captured by the Prisoner's Dilemma. But the revealed preference defense only works if an explanation is empirically adequate (ignoring for now its false behavioral claims about how people reason). And the Prisoner's Dilemma's explanations aren't empirically adequate even in the very cases that are deemed to be its great successes, or so we have argued. In which case, the revealed preference defense fails.

4.4.3 An agenda setter?

Even if the Prisoner's Dilemma does not explain many social phenomena, might it still play other useful roles? We will discuss here two candidates. The first role, mentioned earlier, is *heuristic*. More particularly, the thought is that even if it were not directly explanatory of individual cases, still the Prisoner's Dilemma might serve as an agenda-setter, structuring research. Descriptively speaking, there is much reason to think that this has indeed happened. But normatively speaking, is that desirable? Maybe sometimes. For example, from the beginning the Prisoner's Dilemma was lauded for making so clear how individual and social optimality can diverge. Moreover, it seems convincing that it has been heuristically useful in some individual cases, such as in inspiring frameworks that better explain entrepreneur–venture capitalist relations (Cable and Shane 1997). This would replicate the similar heuristic value that has been claimed for rational choice models elsewhere, for instance in the design of spectrum auctions (Alexandrova 2008, Northcott and Alexandrova 2009).

Nevertheless, overall we think there is reason for much caution. At a micro level, it is all too easy via casual empiricism to claim heuristic value for the Prisoner's Dilemma when in fact there is none. The WWI example illustrates this danger well – there, the Prisoner's Dilemma arguably turned out to be of *negative* heuristic value. On a larger scale, we have seen the gross

disproportion between on one hand the huge size of the Prisoner's Dilemma literature and on the other hand the apparently meager number of explanations of field phenomena that this literature has achieved. Overall, the concentration on theoretical development and laboratory experiments has arguably been a dubious use of intellectual resources.

4.4.4 A normative role?

The second non-explanatory role that the Prisoner's Dilemma might serve is to reveal what is instrumentally rational. Even if it fails to predict what agents actually did, the thought runs, still it might tell us what they *should* have done. For example, given their preferences, two battalions facing each other across WWI trenches would be well advised to cooperate; that is, if the situation is such that they face an indefinitely repeated Prisoner's Dilemma, then it is rational not to defect.

There is an obvious caveat to this defense though, explicit already in its formulation: the normative advice is good only if the situation is indeed accurately described as a Prisoner's Dilemma. Thus a normative perspective offers no escape from the central problem, namely the ubiquitous significance in practice of richer contextual factors unmodeled by the Prisoner's Dilemma.

4.4.5 The aims of science

Why, it might be objected, should the goal of social science be mere causal explanations of particular events? Isn't such an attitude more the province of the historian? Social science should instead be concentrating on systematic knowledge. The Prisoner's Dilemma, this objection concludes, is a laudable example of exactly that – a piece of theory that sheds light over many different cases.

In reply, we certainly agree that regularities or models that explain or that give heuristic value over many different cases are highly desirable. But ones that do neither are not – especially if they use up huge resources along the way. When looking at the details, the Prisoner's Dilemma's explanatory record so far is poor and its heuristic record mixed at best. The only way to get a reliable sense of what theoretical input would actually be useful is via detailed empirical investigations. What useful contribution – whether explanatory, heuristic, or none at all – the Prisoner's Dilemma makes to such investigations cannot be known until they are tried. Therefore resources would be better directed towards that rather than towards yet more theoretical development or laboratory experiments.